

WHAT IS CLAIMED IS:

1 1. A method for throttling incoming requests to a server having a buffer for
2 storing incoming calls request to being processed by the server, the method comprising:
3 receiving incoming requests;
4 determining whether the server is overloaded;
5 if the server is overloaded then storing the incoming requests in the buffer
6 only if a number of incoming requests currently stored in the buffer is less than an
7 acceptance limit; and
8 if the server is not overloaded, then storing the incoming request in the
9 buffer.

1 2. The method according to claim 1, wherein determining whether the server
2 is overloaded is based on at least an incoming request rate and a processing rate of the
3 server.

1 3. The method according to claim 2, wherein the server is overloaded if the
2 incoming request rate exceeds the processing rate of the server.

1 4. The method according to claim 3, wherein the server is overloaded if the
2 incoming request rate exceeds the processing rate of the server for a predetermined
3 amount of time.

1 5. The method according to claim 4, wherein the predetermined amount of
2 time is set based on a desired delay time of the buffer.

1 6. The method according to claim 1, wherein the acceptance limit is less than
2 100% of a capacity of the buffer while the server is overloaded.

1 7. The method according to claim 1, wherein the acceptance limit is 25% of
2 the capacity of the buffer.

1 8. The method according to claim 1, wherein the incoming requests are
2 telephone calls.

1 9. The method according to claim 1, wherein the incoming requests are data
2 transmissions.

1 10. The method according to claim 1, wherein the acceptance limit is
2 determined based on a desired delay time of the buffer.

1 11. An apparatus for throttling incoming requests, comprising:
2 a server having a buffer that stores incoming calls request to being
3 processed by the server; and
4 a controller coupled to the server that determines whether the server is
5 overloaded,
6 if the server is overloaded then storing the incoming requests in the buffer
7 only if a number of incoming requests currently stored in the buffer is less than an
8 acceptance limit, and
9 if the server is not overloaded, then storing the incoming request in the
10 buffer.

1 12. The apparatus according to claim 11, wherein determining whether the
2 server is overloaded is based on at least an incoming request rate and a processing rate of
3 the server.

1 13. The apparatus according to claim 12, wherein the server is overloaded if
2 the incoming request rate exceeds the processing rate of the server.

1 14. The apparatus according to claim 13, wherein the server is overloaded if
2 the incoming request rate exceeds the processing rate of the server for a predetermined
3 amount of time.

1 15. The apparatus according to claim 14, wherein the predetermined amount
2 of time is set based on a desired delay time of the buffer.

1 16. The apparatus according to claim 11, wherein the acceptance limit is less
2 than 100% of a capacity of the buffer.

1 17. The apparatus according to claim 11, wherein the acceptance limit is 25%
2 of the capacity of the buffer.

1 18. The apparatus according to claim 11, wherein the incoming requests are
2 telephone calls.

1 19. The apparatus according to claim 11, wherein the incoming requests are
2 data transmissions.

1 20. The apparatus according to claim 11, wherein the acceptance limit is
2 determined based on a desired delay time of the buffer.